Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging

Hollenstein, Nora ; Aepli, Noëmi

Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging

Nora Hollenstein
University of Zurich
hollenstein@cl.uzh.ch

Noëmi Aepli
University of Zurich
noemi.aepli@uzh.ch

Abstract

Swiss German is a dialect continuum whose dialects are very different from Standard German, the official language of the German part of Switzerland. However, dealing with Swiss German in natural language processing, usually the detour through Standard German is taken. As writing in Swiss German has become more and more popular in recent years, we would like to provide data to serve as a stepping stone to automatically process the dialects. We compiled *NOAH's Corpus of Swiss German Dialects* consisting of various text genres, manually annotated with Part-of-Speech tags. Furthermore, we applied this corpus as training set to a statistical Part-of-Speech tagger and achieved an accuracy of 90.62%.

1 Introduction

Swiss German is not an official language of Switzerland, rather it includes dialects of Standard German, which is one of the four official languages. However, it is different from Standard German in terms of phonetics, lexicon, morphology and syntax. Swiss German is not dividable into a few dialects, in fact it is a dialect continuum with a huge variety. Swiss German is not only a spoken dialect but increasingly used in written form, especially in less formal text types. Often, Swiss German speakers write text messages, emails and blogs in Swiss German. However, in recent years it has become more and more popular and authors are publishing in their own dialect. Nonetheless, there is neither a writing standard nor an official orthography, which increases the variations dramatically due to the fact that people write as they please with their own style.

So far, there are almost no natural language processing (NLP) tools for Swiss German (Scherrer and Owen, 2010). Considering the fact that the major part of communication between Swiss people of the German part is in dialect, we would like to start building NLP tools for Swiss German dialects.

Furthermore, it is an attempt to deal with dialect varieties directly instead of taking the detour through the standard of a language. Speakers of various dialects increasingly communicate through social media in their own varieties. These interactions are relatively easily accessible and could be used as a source of data. However, there is a lack of natural language processing tools for dialects, which need to be developed first in order to process these data automatically.

We start with training a model for a Swiss German Part-of-Speech tagger, which is one of the first steps dealing with the automatic processing of natural language. Based on a part-of-speech tagged corpus, further processes like semantical analysis, syntactical parsing or even applications like machine translation can be conducted.

In order to train a PoS tagger we need a corpus annotated with parts-of-speech. As such data does not exist yet, we compiled *NOAH's Corpus of Swiss German Dialects* containing Swiss German texts of different genres, and annotated it manually. This is an iterative process alternating between running/training a PoS tagger and manually annotating/correcting the output. The corpus we present in this paper consists of 73,616 manually annotated tokens covering many dialect variations of the German-speaking part of Switzerland.

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 85–94,
Dublin, Ireland, August 23 2014.

In the next section, we will mention some related work before we will have a closer look at the Swiss German dialects and its differences to Standard German in section 3. In section 4 we introduce our corpus including the adapted tagset before we present the application of our corpus to the Part-of-Speech tagging task in section 5.

2 Related Work

Most natural language processing applications focus on standardised, written language varieties, but from a methodological as well as a practical point of view, it is interesting to develop NLP methods for variational linguistics. Even though there are no other resources of this size and no studies on PoS tagging for written Swiss German, there have been a few approaches which share some common aspects with our work. While there are some corpora of spoken texts, such as the Archimob project (Dejung et al., 1999) which comprises transcribed interviews, it is difficult to find resources to build a written Swiss German corpus. One of the rare written resources is the *sms4science* project (Dürscheid and Stark, 2011), a collection of text messages in all official languages of Switzerland as well as Swiss German dialects.

Concerning Part-of-Speech tagging for non-standard dialects, there are some approaches addressing linguistic varieties in historical texts, Hinrichs and Zastrow (2012) and Rayson et al. (2007) for German and English respectively. Furthermore, Diab (2009), Habash and Rambow (2009) and Duh and Kirchhoff (2005) worked on PoS tagging for Arabic dialects. The latter developed a minimally supervised PoS tagger for an Egyptian Arabic dialect, which does not have a standard orthography either, without using any dialect-specific tools.

As far as Swiss German NLP goes, there are approaches to dialect identification (Scherrer and Owen, 2010), dialect machine translation (Scherrer, 2012) and morphology generation (Scherrer, 2013).

3 Swiss German

Swiss German belongs to the Alemannic group of dialects, a branch of the Germanic language family. This group can be split into three linguistic divisions; Low, High and Highest Alemannic, each of which contains a few regions of Switzerland. There is no strict border between the Swiss German dialects and the other Alemannic dialects, rather it is referred to as a dialect continuum. Unlike the continuum among Swiss German dialects, there is a strict separation between Swiss German and Standard German. When it comes to the dialects of Swiss German, one can find the concept of diglossia. Diglossia is defined as a situation in which two languages (or two varieties of the same language) are used under different conditions within a language community. In the case of the German language, Standard German is used in Switzerland nearly exclusively in written context while Swiss German is in daily use, mostly in spoken form but also in informal written contexts (Siebenhaar and Wyler, 1997). However, this distinction is becoming more and more blurred. Schools are one of a few environments where Standard German is expected to be used in spoken language. Unlike the situation in other languages, it is standard in Switzerland to use dialect even in formal situations. In Swiss media, both TV and radio, Swiss German is well represented and commonly used.

With the introduction of emails, text messages, blogs and chats, Swiss German is taking over more and more space in written contexts. Nowadays, especially for the younger generations, it is completely normal to write in Swiss German. However, it is not limited to the private communication. In fact, it is even becoming a cult status to write and publish in Swiss German. Many authors, among them for example Lenz (2013), Schobinger (2014) and Kaiser (2012) write books in their dialect, and newspaper agencies publish newspapers in Swiss German, e.g. *Blick am Abend* (Ringier AG, 2013, 2014). Even the Swiss company *Swatch* has published their annual report 2012 in addition to Standard German, French and English also Swiss German (The Swatch Group AG, 2012). This hype does not seem to cease, in the contrary. Speaking a certain dialect is part of the identification. Swiss are proud of their dialect, which makes it possible to identify their home region if they move to another canton. Despite the big differences, speakers of various dialects usually understand each other, except a few German varieties of the canton Valais which others usually have troubles understanding (Keller, 1961).

3.1 Differences to Standard German

Swiss German differs from Standard German in many aspects such as phonetics, lexicon, morphology and syntax. One of the most significant differences is the vocabulary, which even introduces a new word class not in use in Standard German (see section 4.2). In Swiss German, the Standard German words are sometimes used in a different manner. For instance, in some cases the genus may change: the word *Radio* (radio) as a masculine word (in Swiss German) instead of neutral (in Standard German). However, there are not merely differences between Swiss and Standard German, but also between the different dialectal regions. Scherrer (2011) differs between variations which apply for the whole Swiss German speaking area and differences which appear only in certain dialects and not outside of Allemanic dialects. The differences between the dialects are partly due to the influence from other languages. For instance dialects closer to the French speaking part of Switzerland use different grammatical constructions than Eastern Swiss dialects. In this section we describe some examples of disparities between the Swiss German dialects and Standard German.

In Swiss German there is no preterite tense ("Präteritum") and the pluperfect ("Plusquamperfekt") is used extremely rarely. Both of them are expressed using the present perfect ("Perfekt") or rather a duplication of it (for an example see table 1). Another difference exists with regards to verb tenses and the use of the auxiliary verbs *sein* (to be) and *haben* (to have). For instance, if you are cold, in Switzerland you would say *Ich ha chalt.*, where *ha* is the first person singular of "to have". However, to express yourself in this situation in Standard German, the auxilary verb "to be" is used: *Mir ist kalt.*

Furthermore, there is more freedom in the order of words of a sentence, especially concerning verbs (for an example see table 1) as well as more possibilities to correctly arrange phrases. The overt specification of the subject is another difference. In Swiss German the subject can be dropped in many cases, the information about the person is then usually given in the conjugation of the verb. In the question *Chunnsch au?* (Swiss German) vs. *Kommst du auch?* (Standard German) (Are you coming too?), the subject *du* is not explicitly expressed in the Swiss German version but only in the second person singular conjugation of the verb.

Regarding nouns, the four cases of Standard German (nominative, accusative, dative and genitive) are not all in use in the dialects (Siebenhaar and Voegeli, 1997). Swiss German speakers generally neither speak nor write in the genitive case, apart from a few exceptions e.g. in the dialect of the canton Valais. The genitive is replaced by a possessive dative or a phrase using prepositions. This means, in order to express the German phrase *die Ohren des Hasen* (the bunny's ears), either the possessive dative *am Haas sini Ohrä* or a preposition *d Ohrä vom Haas* (where *vom* is a fusion of an preposition *von* and an article *dem*) is used. Moreover, nominative and accusative forms only differ in personal pronouns, whereas the dative case, if used, is marked with its own determiner and endings for adjectives and nouns.

There are many phenomena, which are treated differently not only in regards to Standard German but also in different dialects. First of all, the lexicon varies a lot. The variations do not only include different pronunciation but also completely different words. For instance in some regions of Switzerland, the Standard German word *Butter* (butter) is used (even though with a masculine article instead of the feminine one, which is correct in Standard German). In other regions, however, different words such as *Anke* are used instead. Another variation concerns the order of verbs if there is more than one of them in a sentence. It is often inverted compared to Standard German, but this varies according to the dialect. To express a final clause with *um ... zu* (in order to) for instance, people in eastern Switzerland would use the concatenation *zum*. Closer to the French speaking part though, the construction *für ... z* is commonly used, which marks the similarity to the French *pour ...*.

The following sentences in table 1 contain examples of both kinds of differences. On the one hand, there are the Standard German preterite forms *liess* and *hatte*, which are expressed in the perfect tense across dialects: *hat ... (gehen) lassen* and *hat gehabt*. On the other hand, the order of the verbs in the perfect construction (*het gha* vs. *gha hät*) as well as the final clause with *um ... zu* differs from dialect to dialect.

Considering the way people write in Swiss German reveals another characteristic. The aforementioned lack of a spelling standard causes variations not only between different authors but also within texts of

Dialect around Bern	Si **het** ne **la ga**, wü er ne gnue Gäud **het gha**, **für** es Billet **z**'löse.
Dialect around Zurich	Si **hät** ihn **gah lah**, wil er nöd gnueg Gäld **gha hät**, **zum** es Billet löse.
Standard German	Sie **liess** ihn gehen, weil er nicht genug Geld **hatte**, **um** ein Billet **zu kaufen**.
English	She **let** him go because he **did** not **have** enough money **to** buy a ticket.

Table 1: Differences between dialects and Standard German

the same author. As people write how they speak, they are not consistent and may spell the same word differently in the same sentence. They are also free to merge any words, which is quite common. Joining words into compounds is not an unseen phenomena in Standard German either. However, a compound is a word consisting of more than one stem, which can act as one word with one corresponding part-of-speech (usually the one of the last part), e.g. *Skilift* (ski lift). In Swiss German, the process of merging words rather resembles the phenomena of clitics, i.e. phonologically bound to another word (Loos et al., 2004). For example *gömmer* is Swiss German for *gehen wir* (we go). *Gömmer* can not be split into verb and pronoun, as the separate occurrences would be *gönd* (first person plural of to go) and *mir* (we). Thus, such merged words are grammatically different words which, however, are phonologically bound and can not stand alone. One phonological word (realised as one alphabetic string limited by white spaces) can even contain the subject, an object and the finite verb of the sentence (see section 4.2 for an example). This means it can not be assigned to one part-of-speech. In section 4.2 we present how we deal with them in the part-of-speech tagging task.

To strengthen our argumentation for the necessity of a Swiss German PoS tagger we compare our results of the training with our corpus with the performance of a Standard German tagger. We run the German model of the most common tagger for Standard German, the TreeTagger (Schmid, 1995), on our Swiss German test set. The tagger reaches an accuracy of 50.8%, which is significantly lower than the result after the training with our corpus.

As we have shown in this section, the dialects of Swiss German differ in many aspects from Standard German. It is not only a different pronounciation or spelling with some variations in the vocabulary. It also involves syntactic differences and constructions which are ungrammatical when transferred to German. Therefore we argue against a normalisation of Swiss German as a mapping to Standard German, a frequently proposed approach dealing with varieties.

4 Corpus Creation

We compiled a Swiss German dialect corpus in order to provide resources to work with Swiss German. Furthermore, we applied the corpus to the basic natural language processing task of Part-of-Speech tagging as a first application. Therefore, we specified a tagset for Swiss German and annotated the corpus according to this tagset.

4.1 *NOAH's Corpus of Swiss German Dialects*

We present *NOAH's Corpus of Swiss German Dialects*, a unique resource for Swiss German. We compiled a Swiss German corpus containing manually annotated part-of-speech tags of 73,616 tokens. As the first annotated resource for written texts in Swiss German dialects, the goal is to cover various text genres as well as different dialects from all regions of Switzerland. *NOAH's Corpus* is freely available for research.[1]

In *NOAH's Corpus*, we include articles from the Alemannic Wikipedia (Wikipedia, The Free Encyclopedia, 2011) in five major dialects (Aarau, Basel, Bern, Zurich and the Eastern part of Switzerland) and a Swiss German special edition of the newspaper *"Blick am Abend"* (Ringier AG, 2013), which was published in 2013. In addition, we added sections of the Swiss German dialect version of the official annual report of the *Swatch* company from 2012 (The Swatch Group AG, 2012). Furthermore, we incorporated extracts of novels from the Swiss author Viktor Schobinger (Viktor Schobinger, 2013) which are written exclusively in the Zurich dialect. Finally, we also included three blogs from *BlogSpot* in various dialects as a web resource. The detailed token quantities for each text source are shown in table 2.

[1] http://www.cl.uzh.ch/research/downloads.html

Text source	No. of tokens
Alemannic Wikipedia	20,135
Swatch Annual Report 2012	13,386
Novels from Viktor Schobinger	11,165
Newspaper articles	11,259
Blogs	17,671
Total	**73,616**

Table 2: Corpus composition

Manning (2011) suggests that the largest opportunities for improvement in part-of-speech tagging lies in improving the tagset and the accuracy of annotation, even though a perfect annotation of words into discrete lexical categories is not possible because some words do not fall clearly into one category. Thus, since the consistency of annotations in natural language corpora is of great importance for PoS tagging performance, we put great emphasis on the manual annotations. After the annotation of the corpus by native speakers, various consistency checks were conducted. For instance, we checked words with low probabilities in the tagging model and we also conducted random checks for cases of difficult tags.

4.2 Tagset

As the basic tagset we use the Stuttgart-Tübingen-TagSet (STTS), which is the standard for German (Schiller et al., 1999). Because of the differences between German and the Swiss German dialects we additionally introduced the tag *PTKINF* as well as the adding of a "+"-sign to any PoS tag.

The newly introduced tag *PTKINF* represents an infinitive particle suggested by Glaser (2003). It is a commonly used and therefore widely analysed phenomenon for Swiss German dialects with no corresponding word or construction in German. In Swiss German people say *Ich go go poschte.* (I'm going shopping.). The second *go* corresponds to the finite verb *gehen* (to go) in the according Standard German sentence *Ich gehe einkaufen.* The first *go*, however, does not exist in the Standard German version. This particle is probably originally derived from *gehen*. However, as a particle it exceeds the use in *gehen* (Glaser, 2003). This infinitive particle *go* (derived from *gehen*; to go) also comes in other forms like for instance *cho* (derived from *kommen*; to come) and *afa* (probably derived from *anfangen*; to begin). In our corpus we found 37 occurrences of this tag.

Furthermore, we introduce special tags for merged words. Since Swiss German does not have official spelling rules, words can be freely joined. Splitting these words in a pre-processing step would be one approach to deal with them. However, it is not always clear where to split them and would result in strange words as the words phonologically assimilate when merged with others (see section 3.1). Also Manning (2011) suggests that splitting tags seems to be largely a waste of time for the goal of improving PoS tagging numbers.

Instead of splitting, we identify these merged words by using the corresponding STTS-tag for the first part and add a plus sign to show that a given word consists of more than one simple word. There are sequences of words that are commonly joined, but also less common combinations can appear as it depends on the preferences of the writer. A commonly joined sequence is, for instance, *VAFIN+PPER*, a personal pronoun attached to a finite auxiliary verb, e.g. *hets* for German *hat es* (there is). An example for a less commonly joined sequence would be a concatenation of three different parts of speech *VVFIN+PIS+PPER* such as *bruchtmese* for the German words *braucht man sie* (one uses/needs it). Figure 3 shows some more examples of the most frequent combinations (e.g. a verb, a conjunction or a particle followed by a pronoun). We found 1008 occurrences of merged words, which represent 1.37% of all tokens in the corpus.

The STTS-tagset already contains one tag that is a combination of two, namely the *APPRART*, consisting of a preposition *APPR* and an article *ART*. This is used for words like *beim*, which is composed of *bei* and *dem*. However, these are "normal" Standard German prepositions. This is not the case with the word combinations in Swiss German writing habits, where any words of completely different parts-of-speech can be merged together. Using the approach of simply joining the corresponding part-of-speech tags of the words like the *APPRART*-case, we would end up with an infinite tagset. Thus, the approach

PoS tag	Swiss German	Standard German	English
VAFIN+	*isches*	ist es	is it
KOUS+	*dasme*	dass man	that one
VMFIN+	*chame*	kann man	can one
PTKZU+	*zflügä*	zu fliegen	to fly
ADV+	*deetobe*	dort oben	up there

Table 3: PoS tags for compound words

of adding a plus sign allows us to have a clearly defined tagset. Another advantage is that it is possible to identify all the concatenated words easily, looking for PoS tags with a "+"-sign attached. Once the list of all occurrences is given, the corresponding tags can still be modified according to one's requirements for further processing in a text or corpus. Moreover, there is not a huge loss of information due to the omitted part-of-speech information for the other word part(s). For many combinations it is very clear which part of speech follows. Coming across a *PTKZU+* for example, the only possibility for the second part is a verb in the infinitive, a fact that can be inferred from the grammar.

5 Evaluation of PoS Tagging

In order to achieve the best results we trained different statistical, open source PoS taggers: TreeTagger (Schmid, 1995), hunpos tagger (Halácsy et al., 2007), RFTagger (Schmid and Laws, 2008), Wapiti CRF Tagger (Lavergne et al., 2010), TnT (*Trigrams'n'Tags*) tagger (Brants, 2000) and BTagger (Gesmundo and Samardžić, 2012). The BTagger and the TnT tagger reach the best results for our corpus, therefore we did a more detailed evaluation of the tagging results based on these two taggers.

5.1 Results

We evaluated the performance of the BTagger and the TnT tagger over our corpus with 10-fold cross validation. The folds we created are non-stratified, i.e. not contiguous sentences. This is because our corpus consists of diverse kinds of text. If we train the tagger on the whole corpus with diverse kinds of text and then evaluate only on blogs for instance, we will not get a fair result. Thus, in order to get balanced test sets, we chose the sentence for the 10 folds randomly. With the whole corpus as training set, we reach an accuracy of 90.62% with the BTagger and 90.14% with the TnT tagger (see table 4). Considering the 26.36% unknown tokens in average over all test sets, the accuracy for the unknown tokens is surprisingly high.

Accuracy	BTagger	TnT tagger
Unknown tokens	77.99%	72.39%
Known tokens	93.34%	93.26%
Overall	**90.62%**	90.14%

Table 4: Accuracy of taggers over the whole corpus

As stated in section 4.1, our corpus contains texts from different genres. Therefore we additionally evaluated the different text genres individually. The results are shown in table 5. The Wikipedia articles score best with 90.92% accuracy. This is due to the fact that it is the biggest part of the corpus with 20,135 tokens (one third). In addition, the amount of unknown words is not as high as in other texts because the variety of different words is limited to one topic per article. The literary texts are on the second place. This corpus part is only half of the size of the Wikipedia articles. However, the texts are all extracted from the criminal novels of Viktor Schobinger. This means, they are written in one dialect by one person, which reduces the number of orthographic varieties and thus the number of unknown tokens. As table 5 shows, the novels have only 16% of unknown tokens, less than all the other parts.

Furthermore, we analysed the relation between the size of the corpus and the accuracy we achieved (see figure 1). In the case of Swiss German we found that the accuracy increases significantly until approximately 40,000 tokens. Increasing the size of the corpus beyond this amount of tokens is helpful

Text type	Accuracy overall	Accuracy unknown tokens	Accuracy known tokens	Number of unknown tokens
Wikipedia articles	90.92%	75.64%	94.60%	22.7%
Literary texts (novels)	89.37%	70.41%	92.89%	16.0%
Annual report	88.82%	76.95%	92.72%	24.7%
Blogs	88.10%	71.69%	91.73%	18.2%
Newspaper articles	87.17%	71.19%	93.15%	27.4%

Table 5: Results for the different text genres with the BTagger

to cover a larger amount of orthographic varieties and reducing the number of unknown words, but does not considerably improve the accuracy of known tokens.

Another fact that stands out in figure 1 is the difference of the tagger performances for a training set of 10,000 tokens. This is due to the fact that that the BTagger makes use of context information and thus emphasises the transition probability by learning sequences of tags. Therefore, not a huge amount of data is needed to get a comparably good performance (Gesmundo and Samardžić, 2012). The TnT tagger, on the other hand, emphasises the emission probability and does not generalise as well.

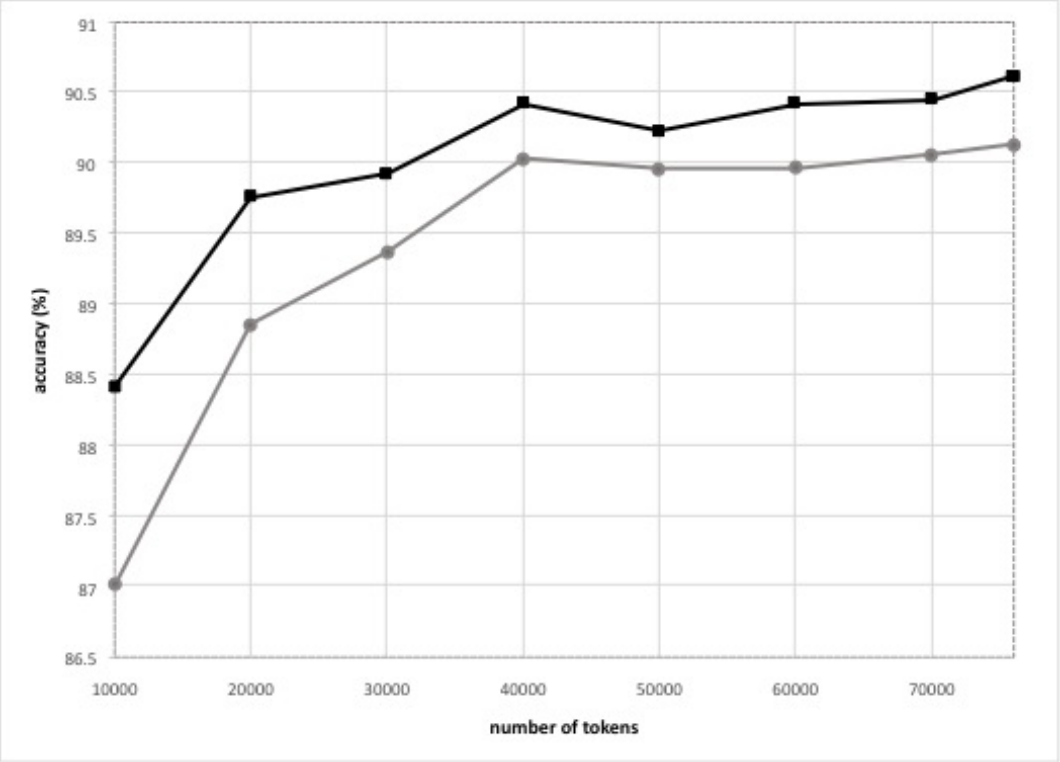

Figure 1: Relation between PoS tagging accuracy and corpus size for the TnT tagger (grey line) and the slightly better results from the BTagger (black line).

In section 3.1, discussing the differences between Standard German and Swiss German, we argue that Standard German tools are not capable of dealing with Swiss German dialects. As an additional experiment we extend our Swiss German corpus with a Standard German corpus to see if the addition of information of Standard German data improves the result. We combined our Swiss German corpus with the *TüBa-D/Z German Treebank* (Telljohann et al., 2006), which contains more than 1,300,000 tokens. The results on a 10-fold cross validation reached an accuracy of 87.6% which is lower than the results for the Swiss German corpus by itself. This implies that the addition of Standard German training data to our Swiss German corpus is not helpful for the training of a Swiss German PoS tagger.

5.2 Error Analysis

The most frequent errors were the confusion of nouns (*NN*) and proper names (*NE*), which represent ca. 15% of all errors. This is also a common problem for Standard German due to the capitalisation of nouns. The different kinds of adjectives and the adverbs as well as various types of verbs are also often mistaken, but these are confusions inside one part-of-speech category. Furthermore, there are many mistakes between articles and some types of pronouns, especially personal and demonstrative. However, this is not surprising as they often have the same form. For example the German indefinite article *ein* is often realised as *es* in Swiss German, the definite article *das* as *s*. The Swiss German *es* also stands for the German neutral personal pronoun *es* if it is not abbreviated to *s*. This issue is exemplified in table 6.

PoS tag	**Swiss German example**	**Standard German**	**English**
ART (definite)	***es** Buech*	**ein** Buch	a book
ART (indefinite)	***s** Buech*	**das** Buch	the book
PPER	***Es** isch rot.*	**Es** ist rot.	It is red.
PPER	***S** rägnet.*	**Es** regnet.	It is raining.

Table 6: Example of the same types with different PoS tags and meanings

5.3 Discussion & Future Work

We achieved reasonable PoS tagging results for the Swiss German dialects considering the low amount of available resources. As stated in section 3, we are dealing with a dialect continuum missing an orthography standard. We neither select one specific dialect (or region) of Switzerland nor do we normalise the data in any way. Thus, our data contains a high amount of hapax legomena, i.e. words which only appear once. This fact explains the considerably lower accuracy for unknown tokens compared to taggers for standardised languages. Furthermore, we include different sources and different text genres in one corpus, which does not simplify the work for a statistical PoS tagger. Thus, it is conceivable that accuracy improvements may be achieved by concentrating on one particular dialect.

In future work we will enlarge *NOAH's Corpus of Swiss German Dialects* by including more texts per dialect in order to reduce the number of unknown tokens. Another approach we are pursuing is to develop a procedure based on lexical distance measures and syntactical patterns in order to map the different orthographic version of a token, so that the tagger can benefit from these mappings. This procedure may also serve as a starting point towards the lemmatisation of Swiss German texts.

The goal of improving Part-of-Speech tagging for Swiss German as well as extending the corpus is to enable and facilitate the development of further NLP tasks, such as dependency parsing, opinion mining or deeper dialectology studies.

6 Conclusion

We have presented our work on compiling a corpus of Swiss German dialects and its application to the training of a Part-of-Speech tagger. As a first resource, our corpus is a stepping stone for natural language processing for the Swiss German dialect area. Training the BTagger on our corpus results in an accuracy of 90.62%. With little post processing effort on the tagger output, a PoS-annotated corpus for Swiss German can be obtained and thus resources extended.

NOAH's Corpus of Swiss German Dialects contains 73,616 tokens from texts of different genres in different dialects, manually annotated with PoS tags. We are happy to share it with interested parties. The corpus including the PoS tags can be downloaded in XML format.

Acknowledgements

We are grateful to the Institute of Computational Linguistics of the University of Zurich for their support. We would like to thank Martin Volk and Simon Clematide for valuable comments and suggestions. Furthermore, many thanks to Tanja Samardžić for inputs concerning the PoS taggers and David Klaper for providing some of the raw data for the corpus.

References

Thorsten Brants. TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics, 2000.

Christof Dejung, Thomas Gull, and Tanja Wirz. *Landigeist und Judenstempel: Erinnerungen einer Generation 19301945*. Limmat Verlag, 1999.

Mona Diab. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, 2009.

Kevin Duh and Katrin Kirchhoff. POS tagging of dialectal Arabic: a minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 55–62. Association for Computational Linguistics, 2005.

Christa Dürscheid and Elisabeth Stark. SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. *Digital Discourse: Language in the New Media*, pages 299–320, 2011.

Andrea Gesmundo and Tanja Samardžić. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 368–372. ACL, 2012.

Elvira Glaser. Schweizerdeutsche Syntax: Phänomene und Entwicklungen. In Beat Dittli, Annelies Häcki Buhofe, and Walter Haas, editors, *Gömmer MiGro?*, pages 39–66, Freiburg, Schweiz, 2003.

Nizar Habash and Owen Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2009.

Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, Prague, Czech Republic, 2007.

Erhard Hinrichs and Thomas Zastrow. Linguistic annotations for a diachronic corpus of German. In *Proceedings of the 10th Workshop on Treebanks and Linguistic Theories*, Heidelberg, 2012.

Renato Kaiser. *UUFPASSÄ, NÖD AAPASSÄ!* Der gesunde Menschenversand, 2012.

R.E. Keller. *German dialects: phonology and morphology, with selected texts*. Manchester University Press, 1961.

Thomas Lavergne, Olivier Cappé, and François Yvon. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Pedro Lenz. *I bi meh aus eine*. Cosmos Verlag AG, 2013.

Eugene Loos, Susan Anderson, Day Dwight, Paul Jordan, and Douglas Wingate. *Glossary of linguistic terms*. http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsACliticGrammar.htm, 2004.

Christopher D. Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *Computational Linguistics and Intelligent Text Processing*, pages 171–189, 2011.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. 2007.

Ringier AG. Blick am Abig. http://epaper.blick.ch/webreader/baa/download/?doc=BAA280513ZH, May 2013.

Ringier AG. Blick am Abig. http://epaper.blick.ch/webreader/baa/download/?doc=BAA020614ZH, June 2014.

Yves Scherrer. Syntactic transformations for Swiss German dialects. In *First Workshop on Algorithms and Resources for Modelling of DIalects and Language Vareities*, Edinburgh, 2011. EMNLP.

Yves Scherrer. Machine translation into multiple dialects: The example of Swiss German. *7th SIDG Congress - Dialect 2.0*, 2012.

Yves Scherrer. Continuous variation in computational morphology - the example of Swiss German. In *TheoreticAl and Computational MOrphology: New Trends and Synergies (TACMO)*, Genève, Suisse, 2013. 19th International Congress of Linguists. URL `http://hal.inria.fr/hal-00851251`.

Yves Scherrer and Rambow Owen. Natural Language Processing for the Swiss German Dialect Area. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 93–102, Saarbrücken, Germany, 2010.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Guidelines für das Taging deutscher Textkorpora mit STTS, August 1999.

Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995.

Helmut Schmid and Florian Laws. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. *COLING*, 2008.

Viktor Schobinger. *Der Ääschmen und de schtùùrzmord*. Schobinger-Verlaag, 2014.

Beat Siebenhaar and Walter Voegeli. 6 Mundart und Hochdeutsch im Vergleich. In *Mundart und Hochdeutsch im Unterricht. Orientierungshilfen für Lehrer*, 1997.

Beat Siebenhaar and Alfred Wyler. *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. 1997.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Universität Tübingen, 2006.

The Swatch Group AG. Swatch Group Geschäftsbericht 2012. http://www.swatchgroup.com/de/investor_relations/jahres_und_halbjahresberichte/fruehere_jahres_und_halbjahresberichte, 2012.

Viktor Schobinger. Viktor's züritü(ü)tsch. http://www.zuerituetsch.ch/index.html, 2013.

Wikipedia, The Free Encyclopedia. Alemannic Wikipedia. http://als.wikipedia.org/wiki/Wikipedia:Houptsyte, 2011.